Supplementary data for

Direct Sequencing of RNA with MinION Nanopore: Detecting Mutations based on Associations

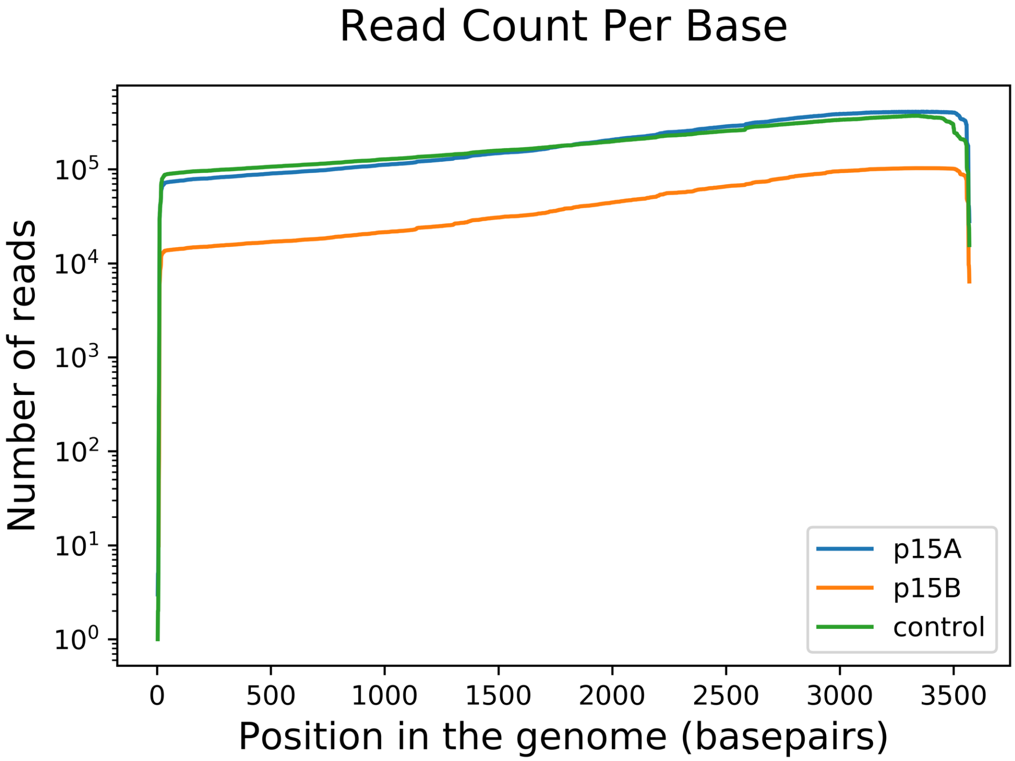


Figure S1. Coverage plots of MinION direct RNA sequencing of MS2 in p15A, p15B and control. The biased slope towards the left highlights the distribution of read lengths, where a large proportion of shorter reads was produced by MinION, and a smaller fraction reached the full length of MS2 (~ 3,569 bp).

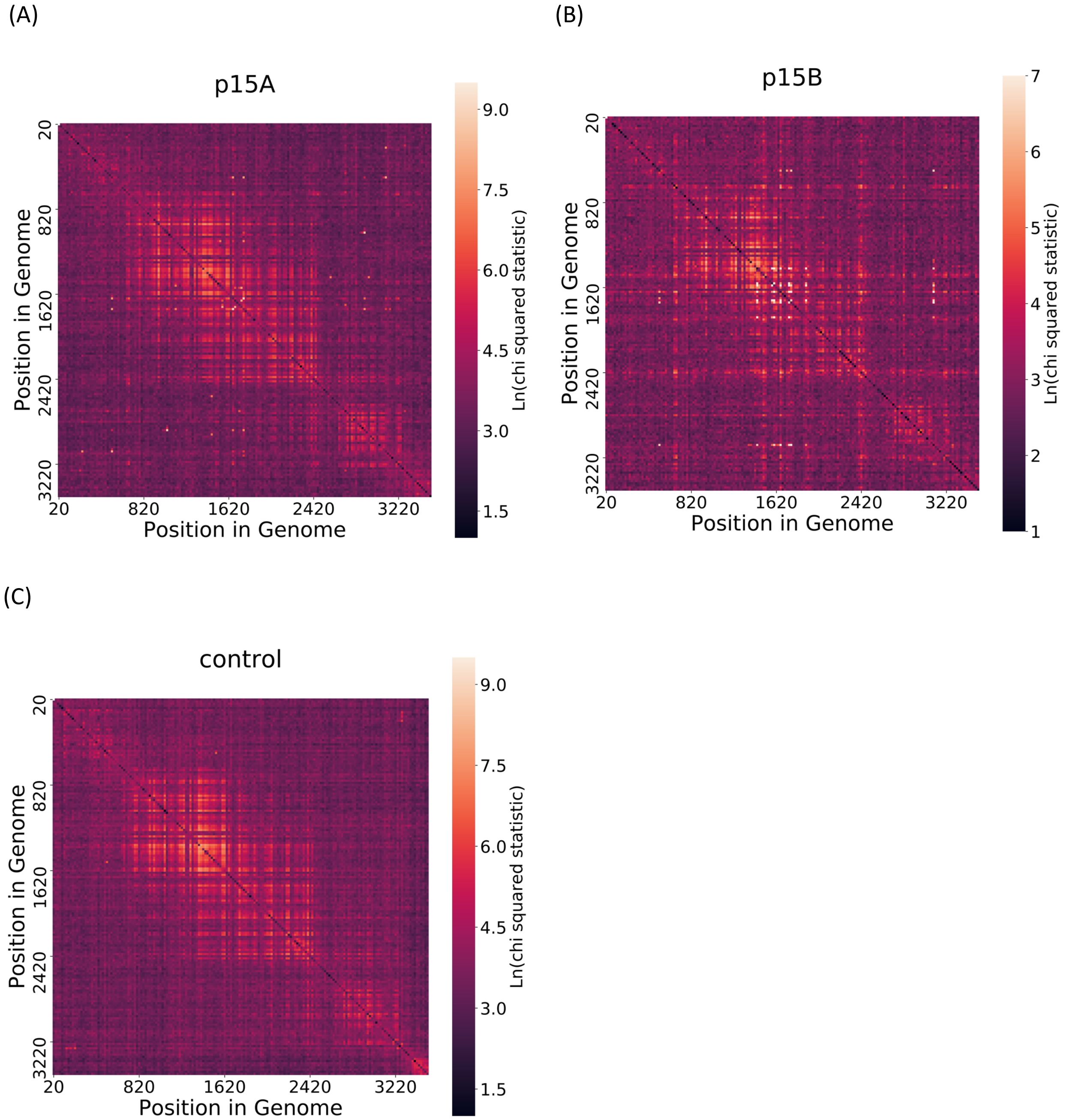


Figure S2. Heatmap of chi square results. The results are presented as the natural logarithm of the chi square statistic. (A), (B) and (C) show the results for p15A, p15B and control respectively. Results are scaled down: each group of 20x20 statistics are presented as the maximum statistic of the group. Statistics for two positions 15 base pairs or less away from each other were removed. Samples show an overall very similar pattern, meaning that MinION sequencing has a tendency towards specific errors for

a specific genome sequenced. The relationships between two real mutations are seen as brighter than their surrounding area.

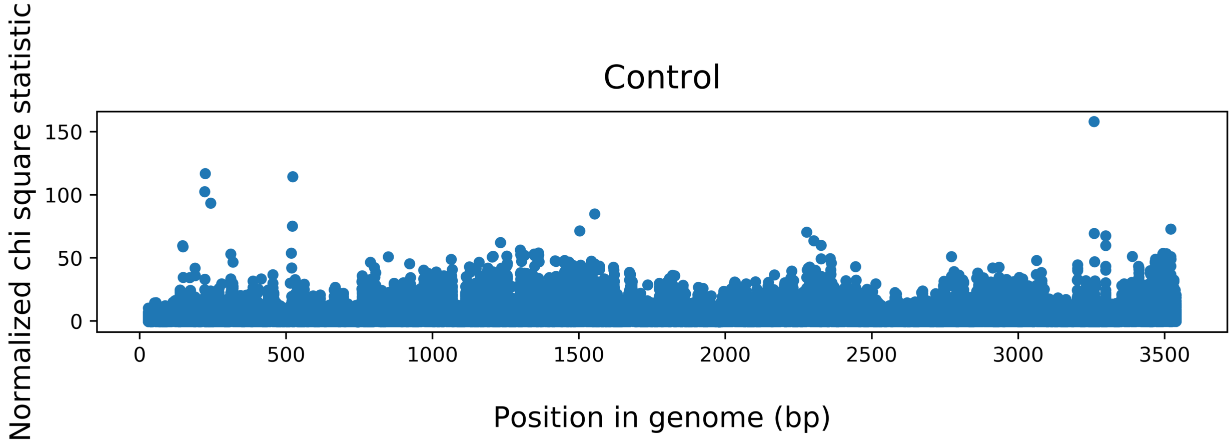


Figure S3. Chi-square statistics plotted along the genome for the control sample minION sequencing, p1A. Details are as in Fig. 4, yet notice the scale difference between the two plots. The control sample associations are much less prominent than the associations for p15A and p15B.

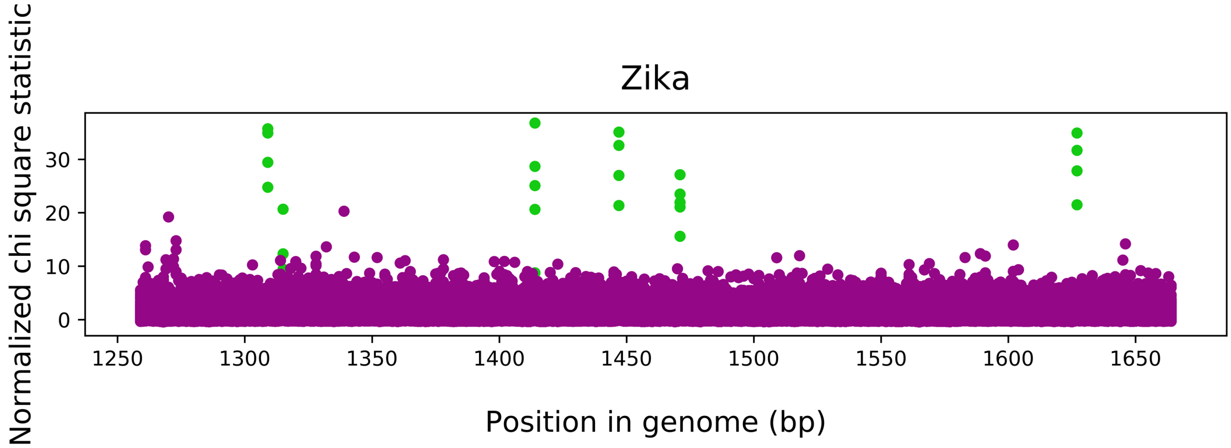


Fig S4. Chi-square statistics plotted along the genome for the Zika virus amplicon of positions 1229-1665. The association between the six positions with true mutations are marked in green. As can be seen, five out of the six mutations have associations higher than any associations between other positions, whereas the associations for the sixth mutation at position 1315 are slightly less significant.

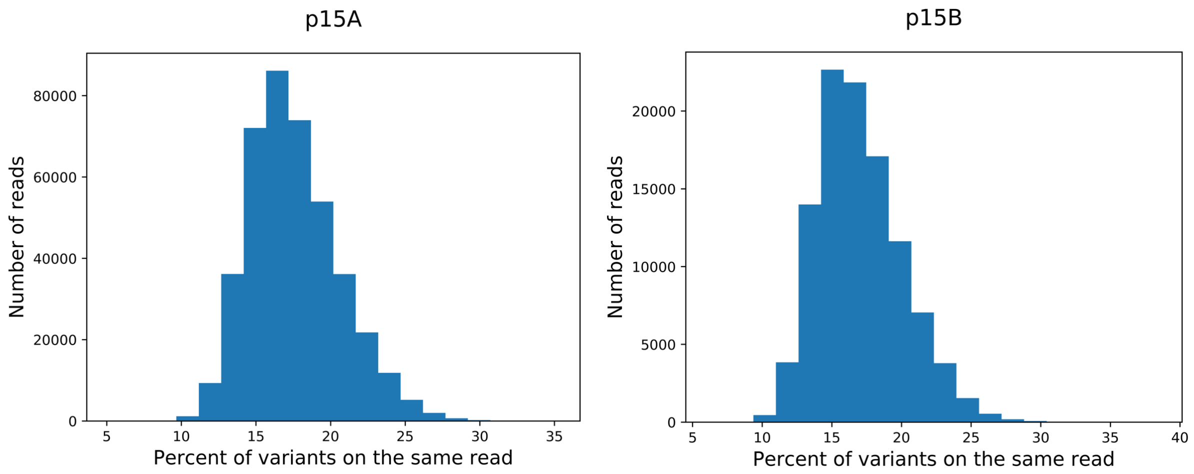
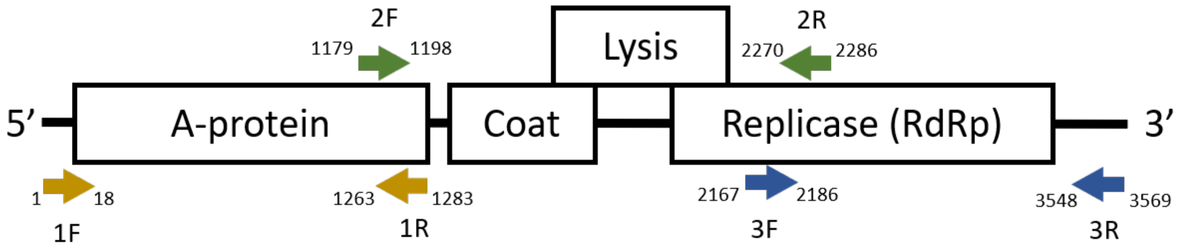


Figure S5. The distribution of the percent of variants observed on the same MinION read.

(A)



(B)

Oligo A: 5'- /5PHOS/GGCTTCTTCTTGCTCTTAGGTAGTAGGTTC

Oligo B 5'- GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGCCTGGGTGGTAACTAGCCAAGCAG

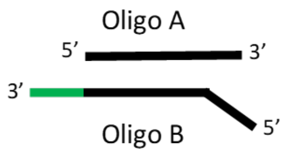


Fig S6. Primers for sequencing. (A) Amplification of the MS2 genome in three amplicons for the Illumina MiSeq library. (B) MinION custom RTA primer, the green sequence is complementary to the 3' end of the MS2 genome.

Table S1. The differences between the MS2 reference (GenBank ID V00642.1) and the consensus sequence from passage 1.

Position	Passage 1 consensus	V00642.1
2002	A	G
2004	C	G
2005	G	A
2006	G	T
2007	T	C
2159	C	T
2160	T	C
2426	C	T
2429	T	C
2591	T	A
3038	C	T
3451	C	-
3452	C	-
3463.01	-	C
3463.02	-	C

Table S2. Pairs of positions having a normalized chi score higher than the normalized chi-square cutoff of 114 defined by the control. The results also include a "local maximum" column, determining whether the normalized chi score of the pair answers the condition of being higher than its surrounding, thus being considered a real mutation by our method. Positions that are identified as false positives when comparing our method to the Illumina results are marked in red.

Position 1	Position 2	Chi score statistic	Normalized chi score statistic	Local maximum
P15A				
1050	2901	6194.947	641.6877	TRUE
1664	1764	9216.712	576.5063	TRUE
2901	1050	6194.947	512.5677	TRUE
1688	1744	3130.478	501.9177	TRUE
1560	1744	2547.602	460.6541	TRUE
531	3100	3206.283	438.3003	TRUE
1764	1664	9216.712	423.9477	TRUE
3100	531	3206.283	402.5826	TRUE
1744	1688	3130.478	387.3511	TRUE
3105	252	1675.803	376.9591	TRUE
535	1764	2385.867	367.655	TRUE
252	3105	1675.803	351.425	TRUE
535	1664	2103.958	324.203	TRUE

1744	1560	2547.602	315.1994	TRUE
1663	1764	1605.107	235.4402	FALSE
1131	1764	1312.509	223.9578	TRUE
2953	1475	1096.916	221.1228	TRUE
2585	1475	1166.538	215.7766	TRUE
1051	2901	1367.668	209.2721	FALSE
1549	1131	1076.928	200.0303	TRUE
1131	1549	1076.928	183.7197	TRUE
2735	1724	615.3027	151.5377	TRUE
2901	1764	1801.675	149.0306	TRUE
1763	1664	909.4032	142.3082	FALSE
1475	2585	1166.538	142.1301	TRUE
1475	2953	1096.916	133.6314	TRUE
1664	535	2103.958	131.567	TRUE
2356	1764	539.1442	126.3444	TRUE
1655	1764	475.5166	119.5437	TRUE
P15B				
3114	1764	4009.217	361.0727	TRUE
1440	1744	5119.376	333.4319	TRUE
1611	1744	5385.559	332.46	TRUE
1764	3114	4009.217	318.1413	TRUE
1744	1611	5385.559	302.9105	TRUE
1744	1440	5119.376	287.9376	TRUE
1440	1611	4388.441	285.8182	TRUE
1611	1440	4388.441	270.8996	TRUE
3113	1764	980.8612	256.0495	FALSE
3112	1764	1038.127	253.3373	FALSE
3114	1664	2803.999	252.5169	TRUE
1441	1744	778.0159	239.5114	FALSE
1664	3114	2803.999	229.2336	TRUE
1664	1764	2671.664	218.4126	TRUE
1906	1744	1782.185	218.3015	TRUE
1764	1664	2671.664	211.9908	TRUE
535	3114	705.8766	205.9131	TRUE
1906	1611	1670.611	204.6286	TRUE
1441	1611	644.9021	198.5045	FALSE
1593	1744	1016.355	197.3718	TRUE
1593	1611	1006.947	195.5436	TRUE
1730	535	300.2778	188.1661	TRUE
3113	1664	715.8328	186.8283	FALSE
1906	1440	1505.701	184.4198	TRUE
3112	1664	745.9047	181.9814	FALSE
535	1764	614.4724	179.2326	TRUE

1593	1440	892.3381	173.2721	TRUE
1592	1611	822.6934	152.2253	FALSE
1592	1744	815.9222	150.971	FALSE
1592	1440	751.9454	139.1195	FALSE
3109	1764	361.04	127.5835	TRUE
1763	3114	318.8337	119	FALSE

Supplementary text

AssociVar Chi Square Tests

The chi-square test of independence tests whether count observations on two variables in a contingency table are independent of each other. In our case, the categorical variables are the nucleotides present in two positions, classified as WT or non-WT for every position. The contingency table contains the read counts for every combination of the two positions. Such a table allows calculating the expected counts based on the marginal cell frequency, and thus calculating the chi-square statistic by comparing observed and expected frequencies. For example, for positions 534 and 1407, the contingency tables for the observed counts can be:

	WT in position 1407	Non-WT in position 1407	Total
WT in position 534	4100	900	5000
Non-WT in position 534	400	100	500
Total	4500	1000	

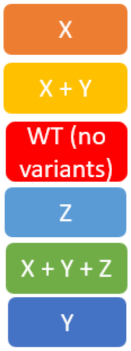
Then the expected counts contingency table will be:

	WT in position 1407	Non-WT in position 1407
WT in position 534	$\frac{4500}{5500} \times \frac{5000}{5500} \times 5500 = 4090.91$	$\frac{1000}{5500} \times \frac{5000}{5500} \times 5500 = 909.09$
Non-WT in position 534	$\frac{4500}{5500} \times \frac{500}{5500} \times 5500 = 409.09$	$\frac{1000}{5500} \times \frac{500}{5500} \times 5500 = 90.91$

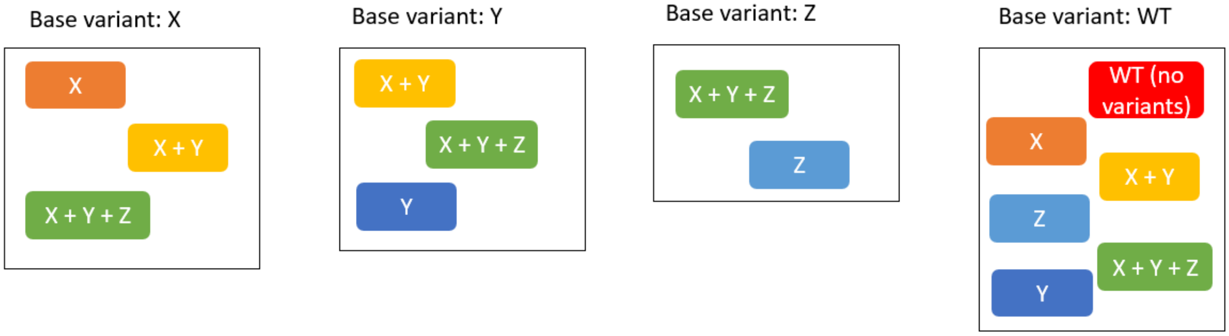
And hence the χ^2 for this pair of positions will be 1.09.

Haplotype/Strain Identification Analysis

A. For variants X, Y and Z, find all possible combinations of variants (haplotypes) that were observed in the sequencing data.



B. Create a group for every base variant, where each group contains all of the haplotypes that contain that variant. WT is also treated as a base for a group which will include all of the observed haplotypes. Haplotypes will hence be present in more than one group.

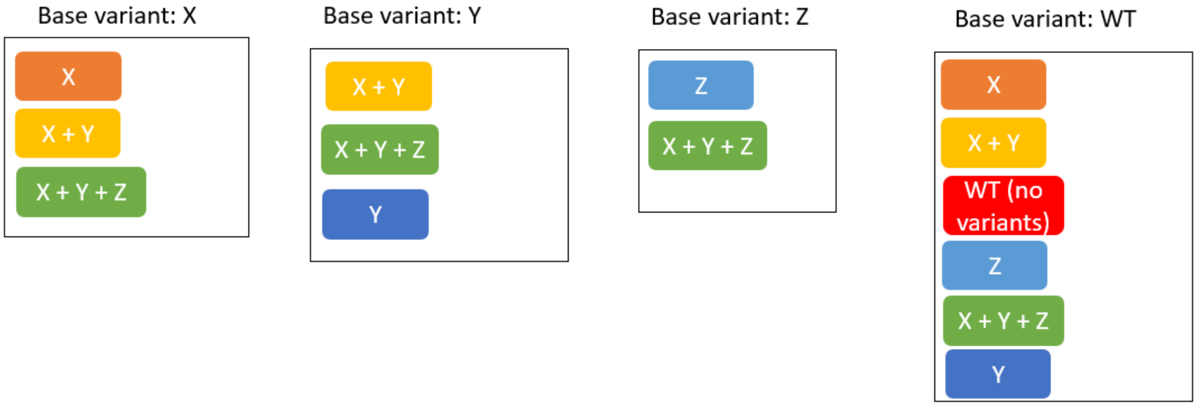

Base variant: X

Base variant: Y

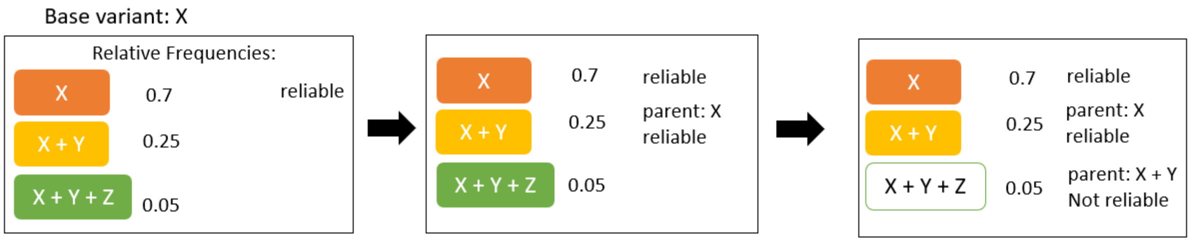
Base variant: Z

Base variant: WT

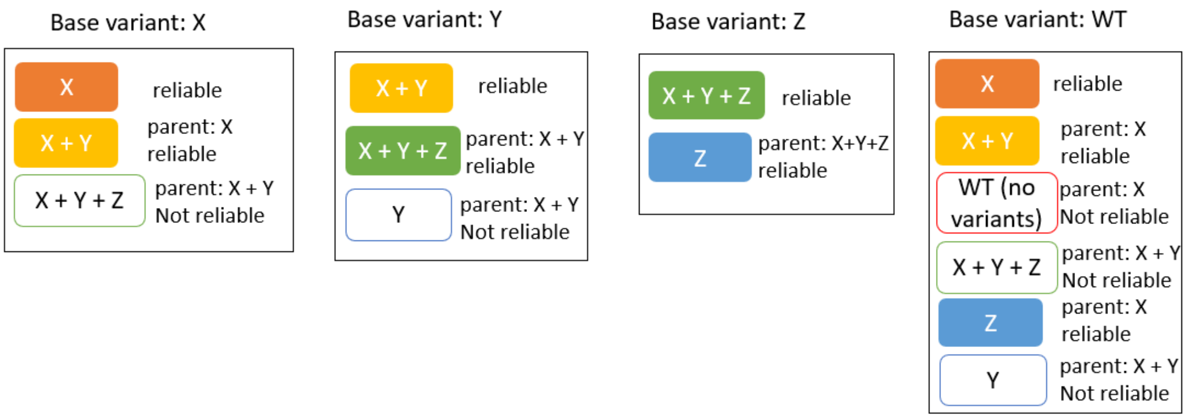
C. Within group, sort by the absolute frequency of the haplotype in the sample, and also calculate the relative frequency of each haplotype in its base variant group.



D. Within each group, iterate through the haplotypes from highest frequency to lowest, classifying each haplotype as reliable or not. The first haplotype is automatically classified as reliable. For every following haplotype, we compare its relative frequency with the probability that it is created by technical errors from the closest haplotype classified as reliable, called its parent haplotype, using the inferred error threshold. For example, if a haplotype has an additional deletion and substitution when compared to its parent haplotype, we require that its relative frequency be higher than the product of 0.214x0.237=0.051 to be classified as reliable (using the 95th percentile error frequencies from Table 1). For example, for base variant X:



We do this within every group:



Base variant: X

X reliable

X + Y parent: X reliable

X + Y + Z parent: X + Y Not reliable

Base variant: Y

X + Y reliable

X + Y + Z parent: X + Y reliable

Y parent: X + Y Not reliable

Base variant: Z

X + Y + Z reliable

Z parent: X+Y+Z reliable

Base variant: WT

X reliable

X + Y parent: X reliable

WT (no variants) parent: X Not reliable

X + Y + Z parent: X + Y Not reliable

Z parent: X reliable

Y parent: X + Y Not reliable

E. For haplotypes appearing in more than one group, it is enough to be classified as reliable in one group to be classified as reliable overall. Finally, we report the overall list of reliable haplotypes and recalculate their relative frequency within the set of reliable haplotypes.

Reliable strains:

	X	X + Y	X + Y + Z	Z
Absolute frequencies:	0.56	0.2	0.04	0.03
Recalculate frequencies within reliable haplotypes:	0.67	0.24	0.05	0.04

Identifying RNA modifications using Tombo

We started by running Oxford Nanopore's Tombo, using both the 5-methylcytosine identification and the de novo modification detection. While 5mC detection searches for a specific modification on cytosine bases, de novo modification is more general and performs a hypothesis test against the canonical model based on the genomic sequence for each position in each read. Both methods return the fraction of reads that were found to be modified per position.

We began by analyzing the control sequence of the enolase II yeast gene. This sample was created synthetically and hence is not expected to have any base modifications (Oxford Nanopore Technical Services, personal communication), and so we used it as a means to test the false positive rate for modification detection. Unfortunately, both methods showed a very high false positive (FP) rate (Fig. S7). In the 5mC method, 50% of positions were identified as being modified in over 10% of reads, 5% of positions were identified as being modified in over 92% of reads and 1% were identified as being modified in over 99% of reads. In the de novo detection the FP rate was even higher, with 50%, 5% and 1% of positions identified as modified in over 43%, 94% and 98% of reads, respectively.

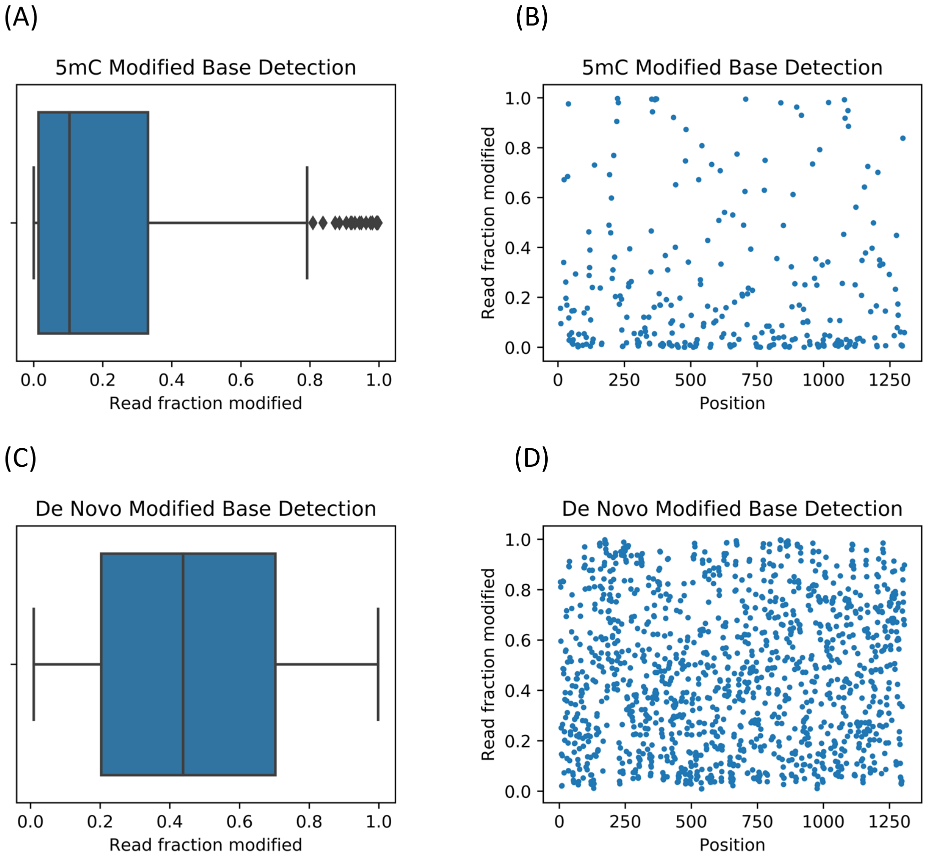


Fig S7. Modification detection using Tombo on a negative control, the enolase II yeast gene. (A) and (C) display the distribution of the read fraction modified per base, (B) and (C) show a visualization of the read fraction modifed per base along the genome. De novo modification detection (C, D) has a higher false positive rate than the 5mC detection (A, B). False positives seem to be distributed evenly along the gene (B, D).

We next ran the same analysis on the MS2 samples. The three MS2 samples show a similar distribution of modification per site to the negative control (Fig. S8). Additionaly, the three samples were highly correlated in terms of modifications per base, as seen in Fig. S9. As the electric signal is compared to the signal expected for the reference sequence, bases containing bona-fide mutations will obviously be identified as having a high modification rate and thus were removed from this analysis (bases containing mutations that presented at over 1% in the MiSeq results). We used a false positive cutoff rate based on the enolase control sample, as suggested previously (*38*), to assess the number of potentially modified sites in the MS2 sample. A 5% cutoff suggested that between 11 and 20 positions underwent 5mC modification in the MS2 samples (between 1% and 2.5% of positions), yet a 1% cutoff suggested that no positions undergo such modification.

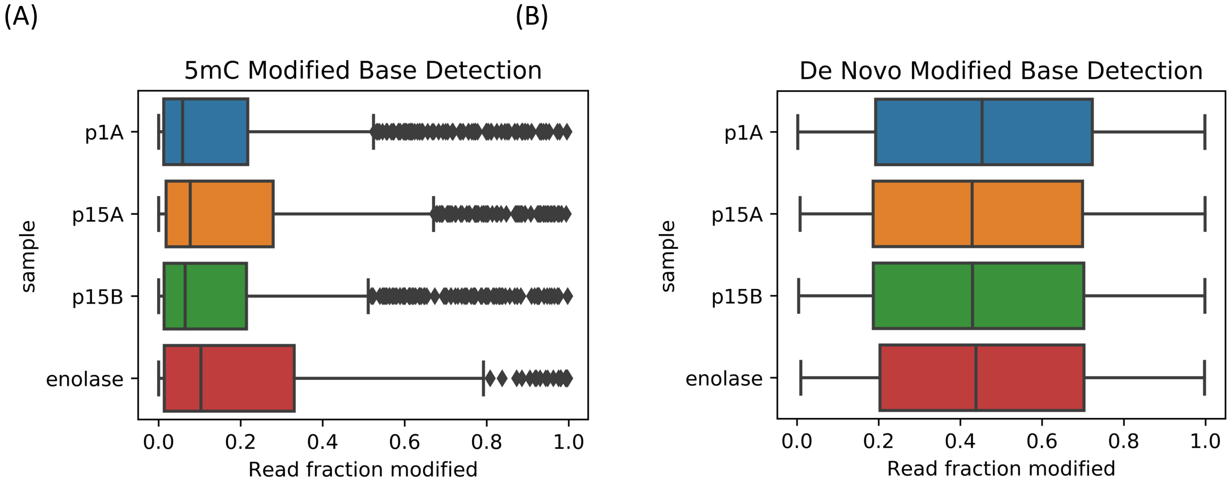


Fig S8. Modification detection using Tombo on samples p1A (control), p15A, p15B and enolase. Plots display the distribution of the read fraction modified per base with 5mC detection (A) and de novo detection (B).

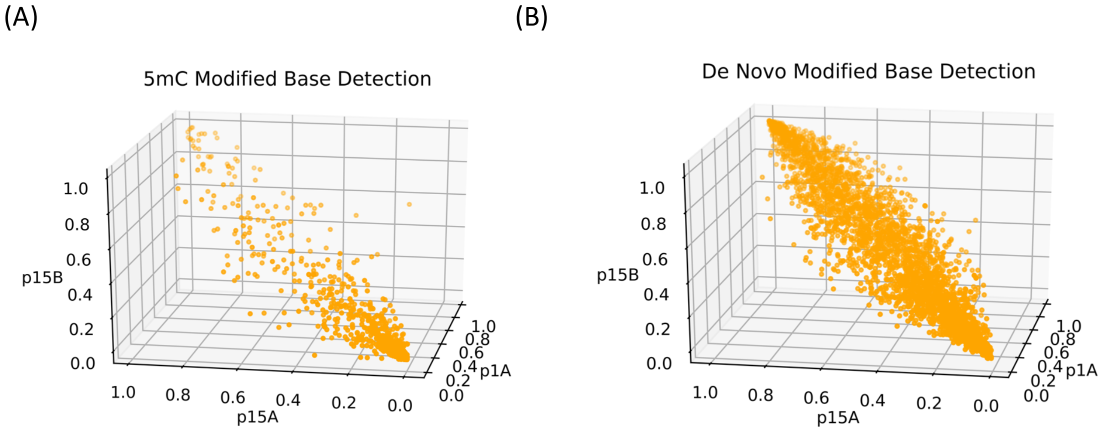


Fig. S9. Comparison between Tombo results for the MS2 samples for 5mC detection (A) and de novo detection (B). Each dot represents a position, and its three coordinates represent the fraction of modification in that position in the three MS2 samples. Samples show a high correlation in the results, yet this could be due to either errors induced by sequence/structural context, or due to modifcations.

We further tested whether inferred modification sites were those that scored highest in our AssociVar score, and found no correlation between the AssociVar score and the fraction of modified reads in either of the MS2 samples (Fig. S10). We conclude that while we cannot currently attribute AssociVar's tendencies to RNA modifications, the presence of RNA modifications and their effect on AssociVar have yet to be ruled out.

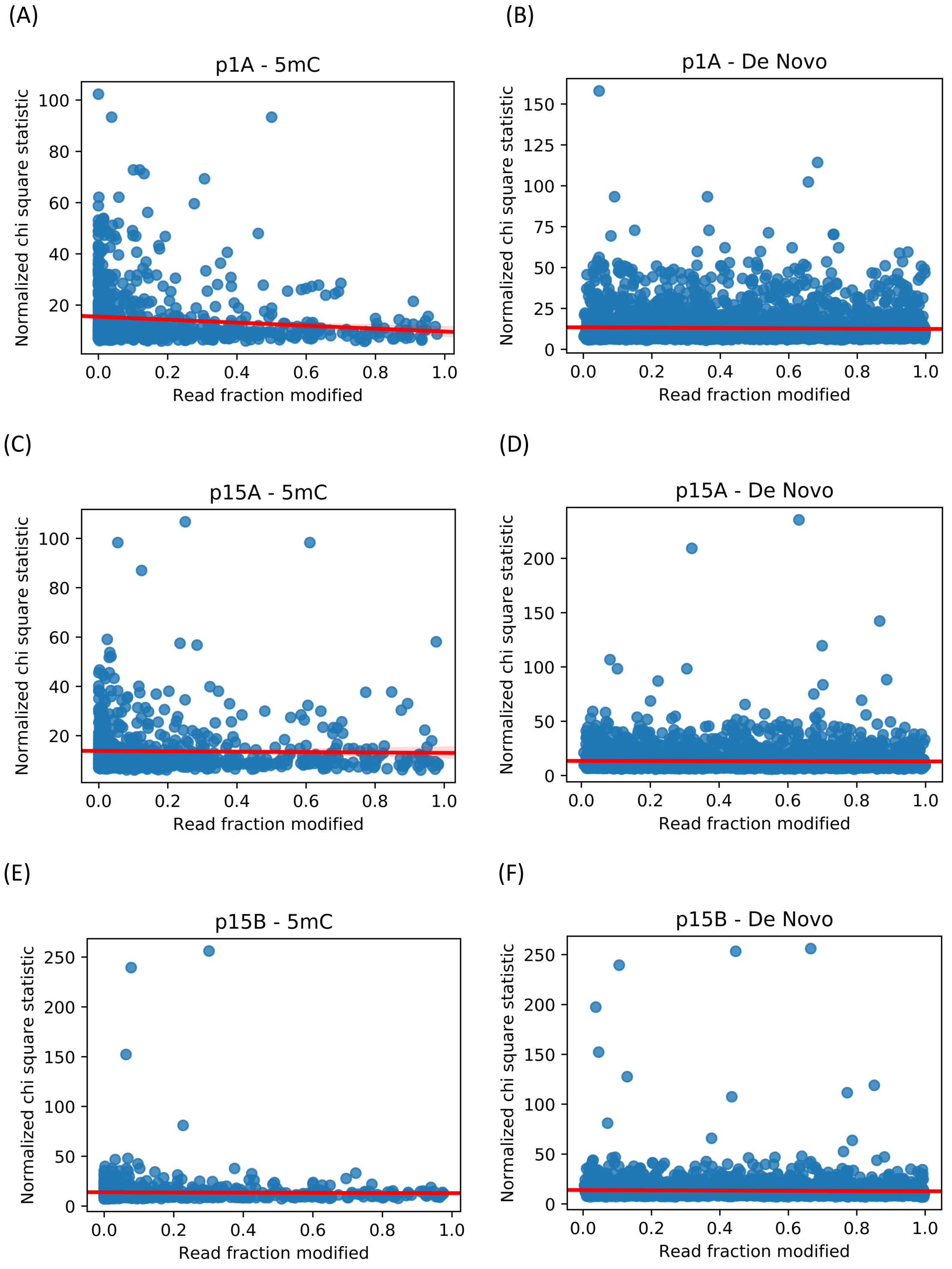


Fig S10. Correlation between Tombo's modified base detection and AssociVar's normalized chi square statistic, calculated for the p1A, p15A and p15B. Every dot represents a position along the genome, and linear regression of the data is plotted as well. (A, C, E) show statistics with Tombo's 5mC identification, and thus only positions where the reference contains a cytosine, while (B, D, F) show de novo modification detection, and thus all positions are plotted. As positions with bona-fide mutations should be picked up by AssociVar and in some degree by Tombo as well, positions with a mutation at a frequency of over 1% according to Illumina MiSeq sequencing were removed from analysis. A regression line is shown in red.